

# DiffusionGemma: Google's Open Model That Writes Text in Blocks, Up to 4x Faster

Kabui, Charles

2026-06-29

---

[Read at ToKnow.ai](#)

---

**DiffusionGemma:**  
**Open Text That**  
**Writes in Blocks**

Google's diffusion model, up to 4x faster text generation

- 4x**  
Faster text generation on dedicated GPUs
- 1,000+**  
Tokens per second on a single H100
- 256**  
Tokens drafted at once, fully in parallel

A 26B open model that activates only 3.8B parameters per step

June 29, 2026

ToKnow.ai

Google released DiffusionGemma, an open model that generates text by diffusion instead of one word at a time. Most language models work like a typewriter, adding one token, a word or word-piece, in sequence. DiffusionGemma drafts a 256-token block at once, refining those placeholder tokens over a few passes, like an image generator sharpening static into a picture. It is a 26-billion-parameter Mixture-of-Experts model that runs only a slice of itself, activating

3.8 billion parameters per step, built on Gemma 4 with a new diffusion head under the Apache 2.0 license. It generates up to 4 times faster on GPUs, over 1,000 tokens per second on one NVIDIA H100.

Compressed to fit within 18GB of video memory, it runs on a single high-end consumer GPU with no cloud bill or per-token fee. Its bi-directional attention, where every token can see every other, suits work that is not strictly left-to-right: in-line code editing, filling gaps in text, or fixing a block's formatting at once. Google notes the catch plainly: output quality sits below standard Gemma 4, so this is built for fast, interactive local work, not top-quality answers.

Open diffusion language models are not new, [LLaDA2.0-Uni](#) took a similar route across text and images, but DiffusionGemma's focus is local speed. Word-by-word generation leaves a single user's GPU idle; drafting in parallel keeps it busy. In high-volume cloud serving, where batching already saturates the hardware, that edge mostly disappears.

Read More: [Google Gemma 4, the open base model DiffusionGemma builds on](#)

Sources:

- [DiffusionGemma: 4x faster text generation \(Google blog\)](#)
- [DiffusionGemma model overview \(Google AI for Developers\)](#)
- [DiffusionGemma \(Google DeepMind\)](#)
- [google/diffusiongemma-26B-A4B-it weights \(Hugging Face\)](#)

---

*Disclaimer: For information only. Accuracy or completeness not guaranteed. Illegal use prohibited. Not professional advice or solicitation. Read more: [/terms-of-service](#)*